The Impact of Telephone Channels on the Accuracy of Automatic Speaker Recognition

Ivan Jokić, Stevan Jokić, Milan Gnjatović, Milan Sečujski, and Vlado Delić, *Member, IEEE*

Abstract — **This paper presents an experimental study on the impact of telephone channels on the accuracy of automatic speaker recognition. Speaker models and the design of the recognizer used in this study are based on Hidden Markov models. In order to simulate telephone-quality speech signals, several experimental conditions were introduced taking two control factors into consideration: the type of the applied codec and the probability of transmission errors. In addition, the impact of echo signals – that are often present in Internet telephony – on the accuracy of automatic speaker recognition systems is considered. Finally, the paper provides a brief overview of several methodologies for the adaptation of the recognizer to the expected environmental conditions that may enhance the robustness of the speaker recognizer.**

Keywords — **Automatic speaker recognition, experimental study, telephone channel, echo in VoIP, adaptation.**

I. Introduction

ONE of the important application domains of the automatic speech recognition is automatic speaker recognition. It involves two major research lines: speaker verification and speaker identification. In general, speaker recognition may be text-independent, when the recognition system does not have any information on the textual content of the given utterances, or text-dependent, when the textual content is *a priori* known.

Speech is a complex acoustic signal resulting from transformations that occur at different levels: semantic, linguistic, articulatory and acoustic [1], [2]. Thus, from the methodological point of view, approaches to speaker recognition may be focused on processing low-level or high-level voice features [3]. Low-level features are related to the acoustic transformations in the vocal tract. Mel-Frequency Cepstral Coefficients (MFCCs) are often used in speech recognition as low-level features (also applied in this work). They represent the spectral envelope of the observed speech signal, and carry the information about the timbre of the speaker. High-level features carry prosodic information (i.e., pitch, dynamics and rhythm). Appropriate selection of feature vectors that describe voice signals of the given speakers is of crucial importance for speaker models. In the speaker recognition process, the system compares the formed models with test speech samples. In practice, the most widely used models are stochastic ones – e.g., Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) – and neural networks [4].

From the application point of view, the existence of various types of telephone systems opens the possibility to access a number of telephone-based services. Often, it is necessary to restrict service access to authorized users. Since the voice is one of the biometric characteristics, speaker recognition may be applied for this purpose. However, the constituent parts of a telephone channel – i.e., the encoder, the transmission system, and the decoder – distort the quality of the original voice signal and, thus, hinder the recognition process.

A considerable amount of effort has been already devoted to the research question of enhancing the robustness of speaker recognition systems for the given acoustic environment and technical factors [5], [6]. It has been suggested that the developmental data conditions should match the expected environmental conditions, e.g., telephone channels. The adaptation of the recognizer is usually achieved by transforming the feature vectors (cf. [7]) or by matching the speaker models (cf. [8]).

This paper presents an experimental study on the impact of telephone channels on the accuracy of automatic speaker recognition, and introduces a methodology for the recognizer adaptation. It is organized as follows. Section II introduces the implemented speaker recognition system. Section III describes the simulation of telephone-quality voice signals used in the experiments. Section IV discusses the experimental results. Finally, Section V discusses the possibilities for adaptation of the recognizer in order to enhance its robustness.

II. Implementation of the Speaker Recognizer

The presented study relies on the speech corpus developed by AlfaNum group at the Faculty of Technical Sciences, University of Novi Sad [9]. For the purposes of training and testing the speaker models, utterances produced by five male and five female speakers were selected. The utterances can be classified in the following groups:

The presented study is performed as part of the project “Development of Dialogue Systems for Serbian and Other South Slavic Languages” (TR32035), funded by the Ministry of Education and Science of the Republic of Serbia. The responsibility for the content of this paper lies with the authors.

Ivan D. Jokić is with the Faculty of Technical Sciences, Univ. of Novi Sad, Serbia (phone: 381-64-3526245; e-mail: ibahjokih@gmail.com).

Stevan D. Jokić is with the Faculty of Technical Sciences, University of Novi Sad, Serbia; (e-mail: stevan.jokic@gmail.com).

Milan J. Gnjatović is with the Faculty of Technical Sciences, University of Novi Sad, Serbia; (e-mail: milangnjatovic@yahoo.com).

Milan S. Sečujski is with the Faculty of Technical Sciences, University of Novi Sad, Serbia; (e-mail: secujski@uns.ac.rs).

Vlado D. Delić is with the Faculty of Technical Sciences, University of Novi Sad, Serbia (phone: 381-21-4852533; e-mail: vdelic@uns.ac.rs).

- *Digits:* Each speaker produced two utterances: “one two three four five” and “six seven eight nine zero”. The mean total duration of audio recordings for a speaker is 12.8s.
- *Names:* Each speaker produced a sentence containing his first name, family name and the required number. The mean duration of audio recording of a speaker is 2.5s.
- *Words:* Each speaker produced a set of eleven preset words. All the sets are disjoint. The mean total duration of audio recordings for a speaker is 70.7s.

For a given speaker, each utterance has been recorded only once. Therefore, the research problem was focused on the text-independent speaker recognition. The GMM was used to introduce a separate model for each speaker and to model pauses in speech.

Training of the models and implementation of the speaker recognizer were performed by the Hidden Markov Models Toolkit (HTK) [10]. The HTK was used to form a GMM, i.e., a HMM that has one emitting state only [11], as presented in Fig. 1. The probabilities of transitions between states, except for the transition from the state 1 to the state 2, were selected arbitrarily.

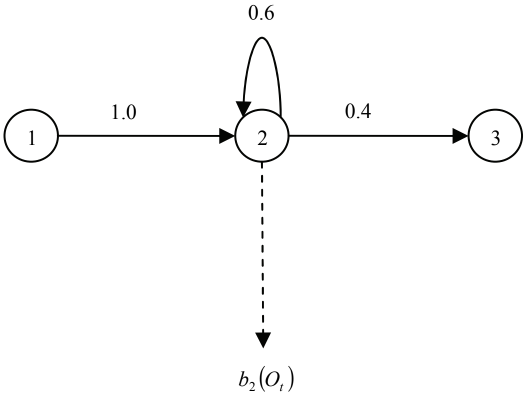

Fig. 1. A speaker model.

The probability distribution of feature vectors for the emitting state is described as follows:

$$b_2(O_t) = \sum_{k=1}^{K} c_{2,k} \cdot \mathrm{N}(O_t, \mu_k, \Sigma_k), \tag{1}$$

where μ_k and Σ_k represent the vector of mean values and the covariance matrix for *k*-th *n*-dimensional Gaussian distribution:

$$\mathrm{N}(O, \mu, \Sigma) = \frac{1}{\sqrt{(2\cdot\pi)^n \cdot |\Sigma|}} \cdot e^{-\frac{1}{2}(O-\mu)^T \cdot \Sigma^{-1} \cdot (O-\mu)}. \tag{2}$$

The same selection parameters were used for the observed models: $K = 64$ и $c_{2,k} = 1/64$ (cf. [12]).

Training of the models and assessment of the recognition accuracy were performed on disjoint sets of utterances from the speech corpus [11]. Feature extraction was performed on speech signal segments by applying a 25ms Hamming window shifted every 10ms. The first 12 MFCCs, zero mel-cepstral coefficient and their first and second derivatives were employed as features. In other words, speaker modeling was conducted by using a mixture of 39-dimensional Gaussian distributions.

III. Simulation of Telephone-Quality Speech Signals

Technical factors such as properties of the telephone channel, bandwidth and transmission characteristics may cause distortion in the output speech signal. Since a speech signal is transferred through digital communication channels, distortion in the signal may be considered as dependent upon the applied codec and the probability of transmission errors. Therefore, in order to simulate telephone-quality speech signals, several experimental conditions were introduced, taking two control factors into consideration: the type of the applied codec and the probability of transmission errors.

In order to manipulate the conditions within Experiment 1, the software library ITU-T STL2005 (ITU-T Software Tool Library 2005) [13] was used to simulate the following codecs:

- **G.711** codec with a bit rate of 64kb/s and a sampling frequency of 8 kHz. This codec is used in the Public Switched Telephone Network (PSTN) and Voice over Internet Protocol (VoIP).
- **G.722** wideband speech codec with a sampling frequency of 16 kHz. Depending on its working mode, this codec supports the bit rates of 64, 56 and 48kb/s. It is used in the Integrated Services Digital Network (ISDN).
- **RPE-LTP** (Regular Pulse Excitation – Long Term Predictor) codec with a bit rate of 13kb/s and a sampling frequency of 8 kHz. This codec is, among others, used in Global System for Mobile Communications (GSM) telephony.
- **G.726** codec with a sampling frequency of 8 kHz that supports the bit rates of 40, 32, 24 and 16kb/s. This codec is used in VoIP telephony.
- **G.727** VoIP codec with a sampling frequency of 8kHz that – with an appropriate choice of core bits, $N_c \in \{2,3,4\}$, and enhancement bits, $N_e \in \{0,1,2,3\}$ – supports the bit rates of 40, 32, 24 and 16kb/s.

In addition, the presented experimental study investigates the influence of a codec that follows the ITU-T G.729 recommendation [14]. In order to simulate its functionality, the electronic file assigned to Annex C+ of this recommendation was applied. This codec is used in VoIP telephony. It uses Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP) with a sampling frequency of 8 kHz. Depending on its working mode, this codec supports the bit rates of 11.8, 8.0 and 6.4kb/s.

In order to manipulate the conditions within Experiment 2 (i.e., transmission errors), the program module *eiddemo.exe* from the software library ITU-T STL2005 was applied to simulate bit errors and frame erasure conditions. The probabilities of bit errors and frame erasure were varied within the intervals $\mathrm{BER} = [0, 0.001]$ and $\mathrm{FER} = [0, 0.1]$, with increments of 0.0001 and 0.01, respectively. The burst factor that describes the sporadic occurrence of these two types of errors was also varied: $\mathrm{BER\_gamma} = \mathrm{FER\_gamma} \in \{0, 0.50, 0.99\}$. These values of the burst factor correspond to totally random, slightly bursty, and totally bursty occurrences of errors, respectively. In most of the experimental settings, the program module for simulation of transmission errors in a telephone channel was applied after the encoder and before the decoder. The only exception is the experimental settings for RPE-LTP codec and G.711-PLC (Packet Loss Concealment) procedure for concealing packet losses. Following the recommendations given in [13], in these experimental settings, simulation of transmission errors was performed after the aforementioned GSM codec and the G.711-PLC procedure. During the simulations of frame erasures in PSTN and ISDN, a 32ms frame was used. The size of the observed frames at the output of the RPE-LTP decoder was 20ms. The size of the observed frames for VoIP channel was 30ms, except for the experimental settings with the G.729 codec, when the frame size was 10ms [11].

IV. Experimental Results

To assess the automatic speaker identification accuracy, the speech signals from the aforementioned speech corpus (cf. also [9]) were used, characterized by a sample rate of 22050Hz and a resolution of 16 bits. After training of the speaker models and testing of the recognizer, the speaker identification accuracy was 100 percent. To assess recognition accuracy on speech signals of telephone quality, in the first experimental phase, the sampling rate of speech signals was decreased in accordance with the bandwidth of the tested telephone channel. Thus, for all the considered telephone channels, except for ISDN with the applied G.722 codec, the sampling rate was decreased to 8kHz. For the G.722 codec, the source signals were downsampled to 16kHz.

After narrowing the spectral range of the source signals, and applying the codecs in the considered telephone channels, the maximum automatic speaker identification accuracy was 90 percent. It was observed at the outputs of the PSTN-G.711, VoIP-G.711, VoIP-G.726 with a bit rate of 40kb/s and VoIP-G.727 with a bit rate of 40kb/s ("4+1"). However, transmission errors were not simulated in this experimental phase [11].

In the second experimental phase, transmission errors were simulated. As expected, increasing the probability of random errors in most cases resulted in the decrease of the recognition accuracy. This is illustrated in Fig. 2. However, the experimental results show that the increased error probability, especially of the bursty errors, does not always imply a decrease of the recognition accuracy [11]. In these cases, errors are probably located in parts of speech signal that are less significant for speaker identification. The identification accuracy at outputs of the PSTN-G.711, ISDN-G.722 with bit rates of 64 and 56kb/s, VoIP-G.711, VoIP-G.726 with a bit rate of 40kb/s, and VoIP-G.727 with a bit rate of 40 kb/s ("4+1") was approximately 80 percent. When telephone channels are considered, the minimum recognition accuracy is observed at the outputs of the VoIP-G.726 and G.727 with a bit rate of 16kb/s. The recognition accuracy at output of the VoIP-

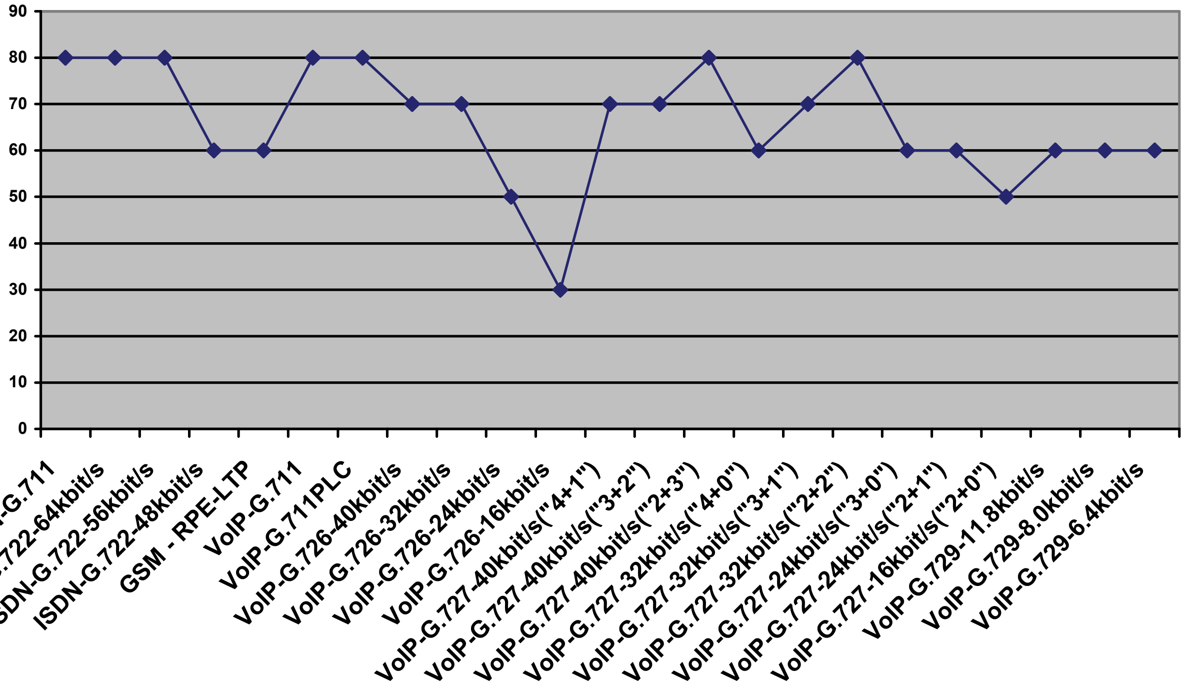

Fig. 2. Speaker identification accuracy depending on the type of telephone channel and the applied codec, where *FER* = 0.05, *BER* = *BER_gamma* = *FER_gamma* = 0.

G.729 was in most experimental settings approximately 60 percent, with rare deviations of ± 10 percent [11].

A. Impact of Echo Signals in VoIP

The time delays of speech frames that occur in VoIP may vary up to 400ms and cause echo at the output of this channel. Thus, the impact of echo signals on the speaker identification accuracy was investigated. In the experimental study, the *Delay/Echo-Simple* effect – a part of the software packet Sony Sound Forge 9.0 – was applied to simulate echo signals at output of the considered VoIP channel. Four time delays of 25, 100, 200 and 400ms were simulated. It is important to note that transmission errors were also simulated in these experimental settings.

Except for a minor number of experiments in which echo was caused using time delays of 25ms, it was observed that occurrences of echo in the channel contribute to better recognition accuracy [11]. Figs. 2 and 3 illustrate the impact of echo signals. Both these figures show recognition results for the same probability of errors in the telephone channels. However, in the experimental settings illustrated in the latter figure, time delays of 100ms were introduced to simulate echo signals at output of the VoIP channel. It is evident that occurrences of echo signals contributed to a more accurate recognition.

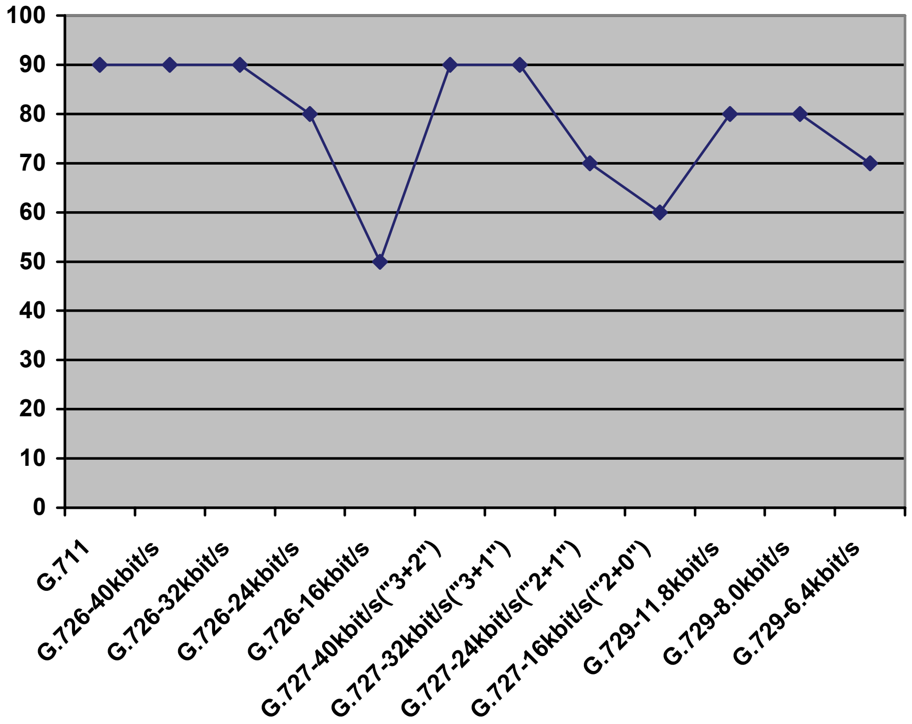

Fig. 3. Speaker identification accuracy at output of the VoIP telephone channel, depending on the applied codec, where *FER* = 0.05, *BER* = *BER_gamma* = *FER_gamma* = 0 and time delay of 100ms.

V. Future Work – Improving the Recognizer Efficiency

One possibility to further improve recognition accuracy is to use larger and phonetically richer speech corpora (cf. [15]). Such approaches aim to overcome negative impacts of telephone channels on the recognition accuracy by better training of speaker models. On the other hand, there are many approaches to improve recognition accuracy by applying appropriate normalization techniques and transformation procedures to feature vectors of speech segments that are the subject of recognition (cf. [16]). These approaches aim to eliminate negative factors that hinder the recognition, i.e., to eliminate negative effects of errors that are present in the channel.

One of the approaches in the literature which seems appropriate with respect to investigation described earlier is C-norm (Cellular normalization). This technique was introduced to compensate channel effects of cellular phones, and performs mapping of channel dependent feature vectors into channel independent feature vectors. The main idea is that the final recognition is performed in a channel independent space. The transformation is defined as follows:

$$x_t^{CI} = f\left(x_t^{CD}\right) = \left(x_t^{CD} - \mu_i^{CD}\right) \cdot \frac{\sigma_i^{CI}}{\sigma_i^{CD}} + \mu_i^{CI}, \qquad (3)$$

where GMM^{CD} and GMM^{CI} denote GMM modeling of the channel dependent feature space and channel independent feature space, respectively, and $i = \arg\max_j \left\{\omega_j^{CD} \cdot p_j^{CD}\left(x_t \middle| \mu_j^{CD}, \sigma_j^{CD}\right)\right\}$ denotes Gaussian distribution within the given GMM^{CD} to which the feature vector x_t^{CD} belongs.

A possible approach to the application requirement that recognition should be as efficient as possible is to reduce decision-making complexity. This reduction may be achieved through relaxation of space in which the recognizer is applied. One of the methods that may address this research problem is principal component analysis (PCA). PCA is introduced to reduce the dimensionality of the observed feature vectors. Its underlying idea is to detect the largest variance directions of the given vectors. Given the covariance matrix of the observed feature vectors, the transformation matrix W can be formed by sorting the eigenvectors in order of the corresponding eigenvalues. The transformed feature is defined as follows:

$$x_t^{PCA} = W \cdot x_t \,. \qquad (4)$$

PCA may be realized by linear algebra, as in (4), or by Multi-Layer Perceptron Artificial Neural Networks (MLPANN).

The aforementioned methods provide appropriate foundation for future work aimed at examining the research questions of improving recognizer efficiency and reducing dependence of the recognizer accuracy on environmental conditions.

VI. Conclusion

The study reported in this paper discusses how automatic speaker identification depends on the applied codec and on transmission errors present in the observed telephone channel. Normally, the codec is *a priori* known for a given telephone channel, and it is possible to take into account its impact when training speaker models. Therefore, a channel-specific approach to designing speaker recognizers is proposed – speaker models should be trained using the speech signal at the codec output.

The accuracy of speaker recognition systems is also discussed with respect to echo signals in VoIP channels. The presented experiments reported better recognition accuracy in cases when echo signals are present in the channel.

Finally, the paper provides a brief overview of some selected methodologies for the adaptation of the recognizer to the expected environmental conditions that may enhance the robustness of the speaker recognizer and improve its efficiency.

References

[1] J. P. Campbell, Jr., “Speaker recognition: a tutorial,” *Proceedings of IEEE*, Vol. 85, No. 9, 1997, pp. 1437-1462.

[2] V. D. Delić, M. S. Sečujski, N. M. Jakovljević, “Action model of human-machine speech communication,” XVI Telekommunications forum *TELFOR 2008*, Serbia, Belgrade, November 25.-27., 2008., pp. 680-683 (in Serbian).

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Margin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” *EURASIP Journal on Applied Signal Processing 2004:4*, 2004, pp. 430-451.

[4] B. R. Wildernoth, “Text-Independent Speaker Recognition Using Source Based Features,” *M. Phil. Thesis, Griffith University Brisbane, Australia*, 2001, pp. 28-37.

[5] T. Kinnunen, H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Communication*, Volume 52, Issue 1, pp. 12-40, January 2010.

[6] S. van Vuuren and H. Hermansky, “On the importance of components of the modulation spectrum for speaker verification,” in *Proceedings 5th International Conference on Spoken Language Processing*, Sydney, Australia, vol. 7, pp. 3205-3208, Nov. 1998.

[7] J. Pelecanos, S. Sridharan, “Feature Warping for Robust Speaker Verification,” ICSA Archive, in *A Speaker Odyssey*, *The Speaker Recognition Workshop*, Crete, Greece, 18-22.6.2001, pp. 213-218.

[8] R. Teunen, B. Shahshahani, L. Heck, “A model-based transformational approach to robust speaker recognition,” *Proc. of ICSLP*, pp. 495-498, 2000.

[9] I. D. Jokić, T. N. Dobrijević, N. M. Jakovljević, V. D. Delić, “Description of speech databases for speaker recognition in Serbian,” XVII Telecommunications forum *TELFOR 2009*, Serbia, Belgrade, November 24.-26., 2009., Proceedings, ISBN 978-86-7466-375-2, pp. 1109-1112 (in Serbian).

[10] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershav, X. (A.) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book (for HTK Version 3.4),” © 1995-1999 *Microsoft Corporation*, © 2001-2009 *Cambridge University Engineering Department*.

[11] I. Jokić, “The Impact of the Telephone Channels to Automatic Speaker Recognition,” *Magister Thesis*, *Faculty of Technical Science, University of Novi Sad, 2010* (in Serbian).

[12] I. D. Jokić, V. D. Delić, N. M. Jakovljević, M. M. Dobrović and S. D. Jokić, “Accuracy of Automatic Speaker Recognition for Telephone Speech Signal Quality,” in *Proc. 8th International Symposium on Intelligent Systems and Informatics*, SISY2010, Subotica, Serbia, ISBN: 978-1-4244-7395-3, pp. 579-582.

[13] ITU-T User’s Group on Software Tools, “ITU-T Software Tool Library 2005 User’s Manual,” *Geneva, August 2005*.

[14] ITU-T Recommendation G.729, “Coding of Speech at 8kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Prediction (CS-ACELP),” *ITU-T Rec. G.729 (01/2007)*.

[15] P. Staroniewicz, “Speaker Recognition for VoIP Transmission Using Gaussian Mixture Models,” *Proceedings of the 4th International Conference on Computer Recognition Systems, CORES’05*, Volume 30/2005, pp. 739-745, DOI: 10.1007/3-540-32390-2_87, May 22-25, 2005, Rydzyna Castle, Poland 2005.

[16] D. Wu, B. Li and H. Jiang, “Normalization and Transformation Techniques for Robust Speaker Recognition,” pp. 311-330, Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert, ISBN 987-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria.